

## GENOTIPAGEM POR ENRIQUECIMENTO E CAPTURA DIRECIONADA DE GENES CANDIDATOS EM *Coffea canephora*<sup>1</sup>

Sinara Oliveira de Aquino<sup>2</sup>; Rémi Tournebize<sup>3</sup>; Pierre Marraccini<sup>4</sup>; Cédric Mariac<sup>5</sup>; Marie Couderc<sup>6</sup>; Kevin Bethune<sup>7</sup>; Philippe Cubry<sup>8</sup>; Alan Carvalho Andrade<sup>9</sup>; Olivier Darracq<sup>10</sup>; Maud Lepelley<sup>11</sup>; Dominique Crouzillat<sup>12</sup>; Catherine Kiwuka<sup>13</sup>; Niels Anten<sup>14</sup>; Stéphanie Manel<sup>15</sup>; Olivier François<sup>16</sup>; Yves Vigouroux<sup>17</sup>; Alexandre de Kochko<sup>18</sup>; Valérie Poncet<sup>19</sup>

<sup>1</sup>Trabalho financiado pela CAPES, EMBRAPA e Fundação Agropolis (França)

<sup>2</sup>Bolsista Pós-doutorado PNPD/Capes, PhD, UFLA, Lavras-MG, Brasil, [saguinobiotec@gmail.com](mailto:saguinobiotec@gmail.com)

<sup>3</sup>Bolsista Pós-doutorado, PhD, Universidade da Califórnia, Berkeley, Estados Unidos, [remi.tournebize@gmail.com](mailto:remi.tournebize@gmail.com)

<sup>4</sup>Pesquisador, PhD, Cirad, UMR IPME, Institute of Agricultural Genetics, Hanoi, Vietnã, [marraccini@cirad.fr](mailto:marraccini@cirad.fr)

<sup>5</sup>Engenheiro, IRD, UMR DIADE, Montpellier, França, [cedric.mariac@ird.fr](mailto:cedric.mariac@ird.fr)

<sup>6</sup>Técnica, IRD, UMR DIADE, Montpellier, França, [marie.couderc@ird.fr](mailto:marie.couderc@ird.fr)

<sup>7</sup>Técnico, IRD, UMR DIADE, Montpellier França, [kevin.bethune@ird.fr](mailto:kevin.bethune@ird.fr)

<sup>8</sup>Pesquisador, IRD, UMR DIADE, Montpellier, França, [philippe.cubry@ird.fr](mailto:philippe.cubry@ird.fr)

<sup>9</sup>Pesquisador, PhD, EMBRAPA Café/ INOVACAFÉ, Lavras-MG, Brasil, [alan.andrade@embrapa.br](mailto:alan.andrade@embrapa.br)

<sup>10</sup>Pesquisador, PhD, Nestlé R&D, Tours, França, [olivier.darracq@rdto.nestle.com](mailto:olivier.darracq@rdto.nestle.com)

<sup>11</sup>Pesquisador, PhD, Nestlé R&D, Tours, França, [maud.lepelley@rdto.nestle.com](mailto:maud.lepelley@rdto.nestle.com)

<sup>12</sup>Pesquisador, PhD, Nestlé R&D, Tours, França, [dominique.crouzillat@rdto.nestle.com](mailto:dominique.crouzillat@rdto.nestle.com)

<sup>13</sup>Doutoranda, NARO, Kampala, Uganda, [catherine.kiwuka@wur.nl](mailto:catherine.kiwuka@wur.nl)

<sup>14</sup>Pesquisador, PhD, Universidade de Wageningen, Wageningen, Holanda, [niels.anten@wur.nl](mailto:niels.anten@wur.nl)

<sup>15</sup>Pesquisador, PhD, EPHE, PSL, UMR CEFÉ, Montpellier, França, [stephanie.manel@cefe.cnrs.fr](mailto:stephanie.manel@cefe.cnrs.fr)

<sup>16</sup>Pesquisador, PhD, TIMC-IMAG, UMR 5525, Grenoble-Alpes, [olivier.francois@univ-grenoble-alpes.fr](mailto:olivier.francois@univ-grenoble-alpes.fr)

<sup>17</sup>Pesquisador, PhD, IRD, UMR DIADE, Montpellier França, [yves.vigouroux@ird.fr](mailto:yves.vigouroux@ird.fr)

<sup>18</sup>Pesquisador, PhD, IRD, UMR DIADE, Montpellier França, [alexandre.dekochko@ird.fr](mailto:alexandre.dekochko@ird.fr)

<sup>19</sup>Pesquisador, PhD, IRD, UMR DIADE, Montpellier França, [valerie.poncet@ird.fr](mailto:valerie.poncet@ird.fr)

**RESUMO:** A captura direcionada de sequências acoplada ao sequenciamento de alto rendimento tornou-se um método poderoso para o estudo da variação de sequências genômicas. Usando o genoma disponível de *C. canephora* (tamanho estimado de 710 Mb/haplóide) como referência, projetamos sondas para 323 genes candidatos selecionados (CGs) visando capturar seus homeólogos em 293 acessos de *C. canephora* da Uganda, 12 clones do Brasil e um subconjunto de 19 indivíduos cobrindo todos os grupos de diversidade de *C. canephora*. Para um total de 324 indivíduos de *C. canephora*, projetamos, com sucesso, cerca de 19.000 sondas e identificamos mais de 16.000 polimorfismos de nucleotídeo único (SNPs) de qualidade, localizados dentro ou próximo a 318 genes anotados, gerando dados genotípicos para populações que estão localizadas em regiões que abrangem diferentes condições climáticas. Nós investigamos questões como desenho das sondas, cobertura do sequenciamento, eficiência da hibridização e captura, e estratégias de análise dos dados por bioinformática.

**PALAVRAS-CHAVE:** Diversidade de *Coffea*, genes candidatos, NGS, enriquecimento de alvo, polimorfismo de nucleotídeo único (SNP).

## GENOTYPING BY TARGET ENRICHMENT AND CAPTURE OF CANDIDATE GENES IN *Coffea canephora*

**ABSTRACT:** Targeted-sequence capture coupled to high-throughput sequencing has become a powerful method for the study of genome-wide sequence variation. Using the available genome of *C. canephora* (estimated size of 710 Mb/haploid) as a reference, we designed probes for the 323 selected candidate genes (CGs) to capture their homeologs in 293 accessions of *C. canephora* from Uganda, 12 clones from Brazil and a subset of 19 individuals covering all the *C. canephora* diversity groups. From 324 individuals of the *C. canephora*, we successfully designed around 19,000 probes and identified >16,000 quality single-nucleotide polymorphisms (SNPs) located inside or in close proximity to 318 annotated genes, generating genotypic data in populations spanning in a range of various climatic conditions. We investigated issues such as probe design, sequencing coverage, evaluation of the hybridization and capture efficiency and strategies of data analysis by bioinformatics.

**KEY WORDS:** *Coffea* diversity, candidate genes, NGS, target enrichment, single nucleotide polymorphism (SNP).

## INTRODUÇÃO

Identificar padrões de variação adaptativa em espécies não modelo, em ambientes naturais, tem sido desafiador devido às dificuldades de se conduzir experimentos e obter medidas diretas, especialmente para espécies de ciclo de vida longo

(Holderegger; Wagner, 2008; Schoville et al., 2012). A maior acessibilidade aos métodos de sequenciamento de alto rendimento e genotipagem facilitou a descoberta de marcadores moleculares em organismos não modelos (Brumfield et al., 2003; Slate et al., 2009; Stapley et al., 2010), incluindo SNPs em genes funcionais ou adaptativos (Cosart et al., 2011).

Notáveis avanços no sequenciamento do genoma completo (WGS - *Whole-Genome Sequencing*) e nas tecnologias de genotipagem proporcionaram uma compreensão mais detalhada sobre os padrões gerais de variação genética (Morin et al., 2004; Pool et al., 2010; Stapley et al., 2010; Tiffin; Ross-Ibarra, 2014) e também sobre a dinâmica populacional ao longo do tempo, levando a uma melhor compreensão do comportamento das espécies diante das mudanças climáticas passadas e futuras.

Os custos decrescentes e o aumento do rendimento de dados das tecnologias de sequenciamento de nova geração (NGS - *Next-Generation Sequencing*) promoveram o desenvolvimento de vários métodos de genotipagem, com base em sequenciamento, permitindo a descoberta simultânea e a genotipagem de números muito grandes de marcadores (Davey et al., 2011). Embora o sequenciamento do genoma completo seja viável, sua aplicação a um grande número de indivíduos em uma única canaleta de sequenciamento de uma tecnologia NGS (por exemplo, MiSeq) permanece tecnicamente desafiadora. De fato, a capacidade de multiplexar indivíduos em uma única corrida é de importância central para a genética de populações, onde o número de amostras é muitas vezes na ordem das centenas. Uma alternativa ao sequenciamento de genomas completos é reduzir a complexidade genômica em uma amostra, visando uma parte do genoma para o enriquecimento seletivo, enquanto se elimina o restante do mesmo. Também conhecido como enriquecimento direcionado ou captura de sequências (Mamanova et al., 2010), esta alternativa tem sido cada vez mais utilizada em uma grande variedade de organismos (Gasc et al., 2016). Os primeiros relatos publicados em humanos (Gnirke et al., 2009; Walsh et al., 2010) e em plantas como milho (Fu et al., 2010), eucalipto (Dasgupta et al., 2015), álamo (Zhou et al., 2012) e pinheiro (Neves et al., 2013) mostraram o potencial dos sistemas de captura para testar variantes de sequências através dos genomas. A captura pode reduzir significativamente os custos em comparação com o WGS pois, apenas locais específicos de interesse são sequenciados com alta cobertura, aumentando a confiança na detecção de variantes reais.

Estas ferramentas genômicas facilitam a identificação de genes com uma função adaptativa com maior confiabilidade do que com métodos tradicionais genéticos ou baseados em observação (Eckert et al., 2010; Hohenlohe et al., 2011; Nielsen et al., 2009), pois inclui genes candidatos (GCs) associados a características de adaptabilidade. Os GCs têm uma função conhecida em um processo particular, uma via metabólica, ou estão relacionados a um fenótipo, e foram encontrados sob seleção em estudos anteriores. Portanto, os GCs têm uma maior probabilidade de estarem sob seleção do que outros genes, especialmente se sua função estiver relacionada às pressões de seleção que variam ao longo da área de estudo (Luikart et al., 2003; Nielsen, 2005).

Explorar a diversidade genética natural de *C. canephora* presente na Uganda (por captura de sequências alvo) pode abrir caminho para a detecção de novas variantes genéticas relacionadas à adaptação à seca. Por exemplo, modificações em sequências codantes de genes de importância conhecida para a tolerância ao estresse abiótico fazem com que esses GCs sejam considerados interessantes para se iniciar novos programas de melhoramento de café com o objetivo de aumentar a resistência desta cultura às mudanças climáticas.

Usando a recente publicação do genoma de *C. canephora* como referência (Denoeud et al., 2014), neste trabalho projetou-se e avaliou-se um conjunto de cerca de 19.000 sondas de captura de sequências para a identificação e genotipagem de quase 16.000 polimorfismos de nucleotídeo único (SNPs - *Single-Nucleotide Polymorphisms*) de qualidade, localizados dentro ou próximos a 318 genes anotados.

## MATERIAL E MÉTODOS

Com base em um conjunto de 323 genes candidatos caracterizados anteriormente como responsivos e com expressão diferencial ao estresse de seca (estudos anteriores realizados pela EMBRAPA/CIRAD/NESTLÉ) e usando a anotação do genoma completo do café (Denoeud et al., 2014), suas sequências foram utilizadas para a síntese das sondas biotiniladas específicas. Para ampliar a região de cobertura pelas sondas, foram também selecionadas regiões flanqueadoras a montante (1 kb) e a jusante (500 pb) de cada gene candidato de modo a incluir potenciais regiões reguladoras. As sondas MYbaits de 120 bp foram projetadas e sintetizadas pela Mycroarray (Ann Arbor, Michigan, EUA). Um total de 21.306 sequências não filtradas foram projetadas. Cada sonda candidata foi mapeada contra o genoma de *C. canephora* (Denoeud et al., 2014) e filtrada com base nos rigorosos critérios do fabricante. O número final de sondas sintetizadas foi de 19.360, cobrindo todos os genes candidatos.

O material vegetal foi composto por folhas de 293 plantas de *C. canephora* de sete florestas selvagens (Budongo, Itwara, Kalangala, Kibale, Mabira, Malabigambo e Zoka), duas estações de pesquisa com material cultivado (Kituza e Kawanda) e materiais de campos que foram abandonados há mais de 50 anos, localizados em Kalangala. Como controles tolerante à seca e suscetível, este estudo incluiu 12 acessos de *C. canephora* Conilon do Brasil, cultivados na estação experimental da Embrapa Cerrados, e por fim, baseado em estudos moleculares anteriores (IRD-Uganda), um subconjunto de 19 indivíduos cobrindo todos os grupos de diversidade de *C. canephora* também foi utilizado.

As extrações de DNA dos 324 indivíduos foram realizadas seguindo o protocolo de Inglis et al. (2018). E, em busca de se garantir o sucesso do enriquecimento por hibridização, todas as amostras passaram por etapas de verificação de

qualidade (gel de agarose) e quantidade (quantificação em espectrofotômetro) antes de serem submetidas ao preparo das bibliotecas.

As bibliotecas foram construídas seguindo os protocolos de Rohland e Reich (2012) e Mariac et al. (2014) e os experimentos podem ser resumidos em 6 etapas, todas elas realizadas individualmente para cada uma das 324 amostras de DNA: 1) fragmentação do DNA; 2) reparação das extremidades não coesivas e fosforilação da extremidade 5'; 3) ligação de adaptadores identificáveis; 4) PCR visando completar falhas existentes e extremidades não coesivas dos adaptadores; 5) PCR em tempo real, buscando a amplificação a partir dos adaptadores inseridos e também a normalização das bibliotecas; 6) e por último, a preparação dos 7 *multiplex* (*pools* com 48 amostras cada) que foram posteriormente hibridizados com as sondas específicas e capturados com partículas magnéticas (kit personalizado MYbaits - Mycroarray).

Os *pools* enriquecidos foram quantificados usando PCR em tempo real e combinados em razões equimolares antes do sequenciamento em uma corrida de Illumina MiSeq (plataforma MiSeq v3, CIRAD, Montpellier, França), que é rápido, simples e permitiu verificar a eficiência da técnica de enriquecimento por hibridização das bibliotecas e também confirmar se o *multiplex* continha todas as amostras em concentrações equimolares, para posteriormente realizar uma corrida de Illumina HiSeq 3000 (Plataforma GeT -PlaGe, GenoToul, Toulouse, França).

As análises das sequências foram realizadas utilizando códigos e scripts da equipe Dynadiv (IRD) e da plataforma *SouthGreen* disponíveis gratuitamente (Mariac et al., 2014; Scarcelli et al., 2016).

A desmultiplexação baseada nos códigos de barras de 6 pb e *index* Illumina foi realizada usando o script DEMULADAPT (PYTHON) disponível gratuitamente, usando um limiar de incompatibilidade de 0. Os adaptadores e bases de baixa qualidade foram removidos usando CUTADAPT 1.8 (Martin, 2011) com as seguintes opções: qualidade de corte = 20, mínima sobreposição = 7 e comprimento mínimo = 35. FASTQC foi utilizado para se avaliar a qualidade das sequências brutas. Os *reads* com uma qualidade média inferior a 30 foram posteriormente descartadas usando um script PERL. Após a desmultiplexagem, remoção dos adaptadores e comparação da qualidade dos *reads forward* e *reverse*, utilizou-se o BWA MEM 0.7.5a-r405 (Li; Durbin, 2009) com a opção -B 4 para o mapeamento dos *reads*, juntamente com as sequências dos GCs selecionados e do genoma do café como referência. Apenas os pares de *reads* mapeados adequadamente (*flag* 0x2), usando o SAMTools 1.1 (Li et al., 2009), foram mantidos nas análises.

## RESULTADOS E DISCUSSÃO

Um total de 19.360 sondas MYbaits® (MP) rigorosamente filtradas foram sintetizadas para *C. canephora*, sendo que estas, cobriram 95% das regiões alvo de 1,3 Mb, representando 0,2 % do genoma total haplóide de *C. canephora*.

Para se garantir um resultado final com alto padrão de qualidade, cada etapa de preparação da biblioteca foi submetida à quantificação e ensaios de controle de qualidade. O DNA de cada indivíduo foi fragmentado usando-se vários ciclos de ultrassom, ajustados a cada amostra individual (média de 8 ciclos), para se obter fragmentos de tamanho médio de 350-450 pb. O tamanho dos fragmentos, obtidos na fragmentação, tem uma grande influência sobre o resultado de um experimento de enriquecimento por hibridização, porque fragmentos menores são capturados por esferas magnéticas revestidas de estreptavidina com maior especificidade do que fragmentos longos.

Para se avaliar a eficiência do enriquecimento na amostra global, foi sequenciado um único indivíduo enriquecido (uma biblioteca) e 48 indivíduos enriquecidos juntos (48 bibliotecas, incluindo a biblioteca do indivíduo enriquecido isoladamente), utilizando-se o sequenciamento Myseq.

Usando as sondas Mybaits, enriquecemos um total de sete *pools* de 48 amostras. O sequenciamento completo foi realizado para uma mistura equimolar dos sete *pools* (48 indivíduos). Primeiramente uma corrida do sequenciamento MiSeq foi realizada para se verificar a proporção de cada um dos sete *pools* na mistura final total.

A execução do MiSeq resultou em uma média de 352.336 *reads* brutas (paired-end) por biblioteca. Após a filtragem de qualidade, remoção dos adaptadores e mapeamento no genoma de referência de *C. canephora*, 316,629 pares de *reads* mapeados adequadamente permaneceram, sendo que 87% com uma cobertura de pelo menos 10x e 54% (191,187 *reads*) dos mesmos, mapeados especificamente nas sequências dos genes candidatos (regiões alvo).

A porcentagem média de *reads* em regiões alvo de bibliotecas construídas por *shotgun* e de bibliotecas enriquecidas foi de 0,26 e 83, respectivamente, confirmando a eficiência da captura em comparação ao sequenciamento completo do genoma, que forneceu apenas 0,26% de *reads* em regiões candidatas (Fig. 1A). Nesse caso, o fator de enriquecimento foi de 320 vezes.

Fatores adicionais relacionados à eficiência e custo são a compatibilidade do método de enriquecimento com abordagens de sequenciamento *multiplex*. Quando 48 bibliotecas foram capturadas, o enriquecimento foi ligeiramente menor do que para uma única biblioteca capturada, mas ainda, eficiente, com um fator de enriquecimento de 70% (Fig. 1B).

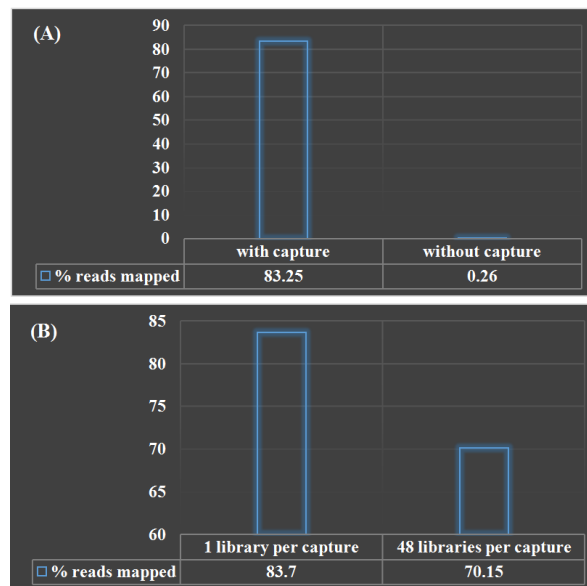


Figura 1. Qualidade do enriquecimento e multiplexação de indivíduos. (A) Porcentagem média de *reads* sob genes candidatos de bibliotecas *shotgun* e enriquecidas. (B) Desempenho da multiplexação de bibliotecas comparando sequências geradas a partir de uma única biblioteca capturada e de um *pool* de 48 bibliotecas.

Confirmamos que cada um dos sete *pools* estavam na mesma proporção no multiplex final por meio da quantidade similar de *reads* totais que cada um dos sete *pools* forneceu (Fig. 2). Além disso, o número de *reads* mapeados também foi semelhante, mostrando que o enriquecimento por captura foi eficiente em cada um dos *pools*.

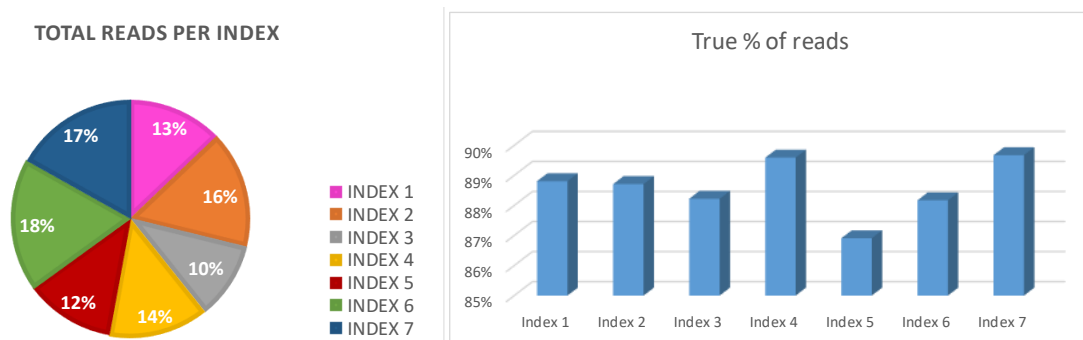


Figura 2. Avaliação da uniformidade do enriquecimento das regiões candidatas. Proporção de *reads* totais de cada um dos sete *pools* no multiplex final sequenciado (figura à esquerda) e porcentagem de *reads* mapeados para um indivíduo aleatório de cada *pool* (figura à direita).

A uniformidade do enriquecimento alvo também foi observada pela distribuição das leituras através dos genes candidatos. Aqui, dois exemplos de genes candidatos, *CcERF034* (codificando uma proteína putativa de 296 aminoácidos homóloga à superfamília AP2/ERF) e *CcDXMT* (enzima envolvida na biossíntese da cafeína) são mostrados abaixo, usando o software IGV (Fig. 3A e 3B).

A captura de genes candidatos pode facilitar a identificação de genes supostamente adaptativos que podem influenciar na aptidão dos organismos e, portanto, serem afetadas pela seleção natural. O conhecimento prévio de genes relacionados à aptidão ou tolerância ao estresse tem se mostrado eficiente para se obter informações sobre as respostas de *C. canephora* às pressões ambientais.

O método descrito aqui, para o enriquecimento de alvos, compartilha alguns aspectos comuns a outras abordagens de redução da complexidade do genoma, como genotipagem por sequenciamento (GBS - *Genotype by Sequencing*) e sequenciamento de DNA associado ao local de restrição (RAD - *Restriction Site Associated DNA*) e suas variações, pois eles também procuram fornecer um grande número de SNPs a um custo menor por amostra, otimizando a profundidade de leitura na região alvo e permitindo se analisar mais amostras por corrida (Silva-Junior et al., 2018).

Como em nossa abordagem, vários trabalhos publicados mostraram o potencial dos sistemas de captura para se testar variantes de sequências em genomas: humanos (Gnirke et al. 2009; Walsh et al., 2010) e plantas como eucaliptos (Dasgupta et al., 2015), pinheiros (Neves et al., 2013) e, mais recentemente, após o desenvolvimento de uma montagem de genoma para a ipê rosa (*Handroanthus impetiginosus*), Silva-Junior et al. (2018) relataram o desenvolvimento de um conjunto de 24.751 sondas de captura para a caracterização e genotipagem de polimorfismos de nucleotídeo único (SNPs) em 18.216 loci distintos, amostrando mais de 10 Mbp do genoma da espécie. Esse sistema identificou cerca de

200.000 SNPs localizados dentro ou próximos a quase 14.000 genes codificadores de proteínas anotados, gerando dados genotípicos de qualidade em populações que abrangem grandes distâncias geográficas em toda a extensão dessa espécie nativa.

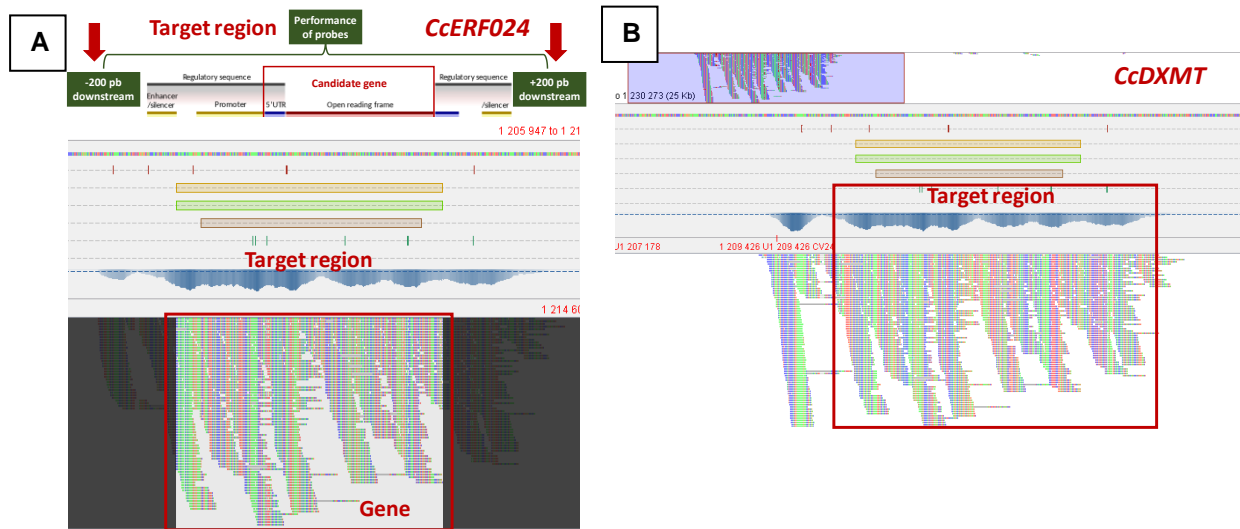


Figura 3. Exemplos da uniformidade do enriquecimento em dois genes candidatos. Acima pode-se observar a distribuição dos *reads* ao longo de dois exemplos de genes candidatos, Fig. 3A. *CcDXMT* (enzima envolvido na biossíntese de cafeína) e Fig. 3B. *CcERF024* (codificador de um fator de transcrição envolvido na resposta ao estresse hídrico).

As vantagens práticas e a crescente acessibilidade dos métodos de captura de sequências alvo também foram revisadas no contexto da genômica evolutiva e ecológica, prevendo uma rápida expansão dessa abordagem (Jones; Good, 2016). Por exemplo, técnicas de enriquecimento direcionadas foram usadas para se aumentar a resolução filogenética dentro do gênero de árvore Neotropical *Inga* (*Leguminosae*) gerando alinhamentos de mais de 0,3 Mbp de sequências codificadoras, revelando quase 5.000 sítios informativos (Nicholls et al., 2015).

## CONCLUSÕES

- 1 - Em comparação com outros métodos de redução da complexidade do genoma, como GBS e RADSeq, o método de captura de sequências alvo, detalhado em nosso estudo atual, tem alto potencial para ser amplamente utilizado na caracterização genotípica de populações de campo e na genética de conservação de *C. canephora*.
- 2 - No nosso caso, utilizamos uma abordagem direcionada com base em genes candidatos específicos (GCs) já identificados por desempenhar uma função nas respostas do café ao estresse abiótico.
- 3 - Relatamos a eficiência da alta multiplexação por captura e provamos que essa estratégia pode ser facilmente estendida ao sequenciamento de *pools* de amostras de DNA.
- 4 - Para uma melhor avaliação da abordagem, também destacamos a importância dos *reads* mapeados nas sequências alvo e o enriquecimento em número de vezes de cobertura em um experimento de captura.
- 5 - Em comparação com outros métodos, a abordagem de captura de GCs usando a técnica MYbaits® tem a vantagem de ser econômica e otimizar a profundidade de leitura na região de destino, permitindo a análise de mais amostras por execução.

## REFERÊNCIAS BIBLIOGRÁFICAS

- BRUMFIELD, R.T. et al. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Evolution and Ecology*, [S.l.], v. 18, n. 5, p. 249-256, 2003.
- COSART, T. et al. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, [S.l.], v. 12, n. 1, p. 347, 2011.
- DAVEY, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, [S.l.], v. 12, n. 7, p. 499-510, 2011.
- DASGUPTA, M. G. et al. Development of genetic markers in eucalyptus species by target enrichment and exome sequencing. *PLoS ONE*, [S.l.], v. 10, n. 1, p. e0116528, 2015.
- DENOEUDE, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, [S.l.], v. 345, n. 6201, p. 1181-4, Sep 5 2014.
- ECKERT, A. J. et al. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology*, [S.l.], v. 19, n. 17, p. 3789-3805, 2010.

- FU, Y. et al. Repeat subtraction-mediated sequence capture from a complex genome. *The Plant Journal*, [S.l.], v. 62, n. 5, p. 898-909, 2010.
- GASC, C.; PEYRETAILLADE, E.; PEYRET, P. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, [S.l.], v. 44, n. 10, p. 4504-4518, 2016.
- GNIRKE, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, [S.l.], v. 27, n. 2, p. 182-189, 2009.
- HOHENLOHE, P. A. et al. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, [S.l.], v. 11, n. 1, p. 117-122, 2011.
- HOLDEREGGER, R.; WAGNER, H. H. Landscape Genetics. *BioScience*, [S.l.], v. 58, n. 3, p. 199-207, 2008.
- JONES, M. R.; GOOD, J. M. Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, [S.l.], v. 25, n. 1, p. 185-202, 2016.
- INGLIS, P. W. et al. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLoS ONE*, [S.l.], v. 13, n. 10, p. e0206085, 2018.
- LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, [S.l.], v. 25, n. 14, p. 1754-1760, 2009.
- LI, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, [S.l.], v. 25, n. 16, p. 2078-2079, 2009.
- LUIKART, G. et al. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, [S.l.], v. 4, p. 981, 2003.
- MAMANOVA, L. et al. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, [S.l.], v. 7, n. 2, p. 111-118, 2010.
- MARIAC, C. et al. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources*, [S.l.], v. 14, n. 6, p. 1103-1113, 2014.
- MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, [S.l.], v. 17, p. 10-12, 2011.
- MORIN, P. A. et al. SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, [S.l.], v. 19, n. 4, p. 208-216, 2004.
- NEVES, L. G. et al. Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant Journal*, [S.l.], v. 75, n. 1, p. 146-156, 2013.
- NICHOLLS, J. et al. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science*, [S.l.], v. 6, n. 710, 2015.
- NIELSEN, R. Molecular signatures of natural selection. *Annual Review of Genetics*, [S.l.], v. 39, n. 1, p. 197-218, 2005.
- NIELSEN, E. E. et al. Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*, [S.l.], v. 18, n. 15, p. 3128-3150, 2009.
- POOL, J. E. et al. Population genetic inference from genomic sequence variation. *Genome Research*, [S.l.], v. 20, n. 3, p. 291-300, 2010.
- ROHLAND, N.; REICH, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, [S.l.], v. 22, n. 5, p. 939-946, 2012.
- SCHOVILLE, S. D. et al. Adaptive genetic variation on the landscape: Methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, [S.l.], v. 43, n. 1, p. 23-43, 2012.
- SILVA-JUNIOR, O. et al. Design and evaluation of a sequence capture system for genome-wide SNP genotyping in highly heterozygous plant genomes: a case study with a keystone Neotropical hardwood tree genome. *DNA Research*, 2018, [S.l.], v. 0, n. 0, 1-11, 2018.
- SLATE, J. et al. Gene mapping in the wild with SNPs: Guidelines and future directions. *Genetica*, [S.l.], v. 136, p. 97-107, 2008.
- STAPLEY, J. et al. Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, [S.l.], v. 25, n. 12, p. 705-712, 2010.
- TIFFIN, P.; ROSS-IBARRA, J. Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, [S.l.], v. 29, n. 12, p. 673-680, 2014.
- WALSH, T. et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proceedings of the National Academy of Sciences USA*, [S.l.], v. 107, n. 28, p. 12629-12633, 2010.
- ZHOU, L.; HOLLIDAY, J. A. Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics*, [S.l.], v. 13, n. 1, p. 703, 2012.